



# EdiFungi: Navigating the Wilderness

An Exploration into Mushroom Foraging, Navigation, and Survival Strategies

# Introduction

Confronting Coronary Heart Disease

## Background:

- Mushrooms, a diverse group of fungi, pose a unique challenge in distinguishing between edible and poisonous varieties due to their varied characteristics.

## Importance of Edibility Classification:

- Accurate prediction of mushroom edibility is paramount for ensuring the safety of individuals engaged in mushroom-related activities. Mistaking a poisonous variety for an edible one can lead to severe health consequences.

## Role of Predictive Modeling:

- Enter the realm of predictive modeling – an innovative approach leveraging machine learning. By harnessing the power of these models, we aim to provide a robust tool for classifying mushrooms, offering a reliable means to differentiate between safe and potentially harmful varieties.

# Objectives

---

## Primary Goals:

- Our primary goal is crystal clear – to develop an advanced predictive model for classifying mushrooms based on their edibility. This model aims to enhance safety and decision-making in mushroom-related activities.

## Importance of Accurate Classification:

- Misclassification can have profound consequences. Accurate predictions are not just a goal; they are a necessity. We'll delve into the potential risks associated with misidentifying mushrooms and highlight the immense benefits of precision in edibility classification.

## Relevance to Mushroom Enthusiasts:

- For mushroom enthusiasts and foragers, our model becomes a reliable companion. By addressing their needs for quick and accurate identification, we empower individuals in making informed decisions during mushroom foraging, ultimately ensuring their safety and well-being.

# Business Problem

The CHD Challenge

## Safety Concerns in Mushroom Foraging:

- Venturing into the world of mushroom foraging is an exciting endeavor, but it comes with inherent risks. The possibility of accidental ingestion of poisonous mushrooms looms, presenting serious safety concerns. It's crucial to address these risks to ensure a secure and enjoyable foraging experience.

## Need for a Reliable Classification Model:

- In the quest for a solution, the need for a dependable model becomes apparent. A model that can swiftly and accurately distinguish between edible and poisonous mushrooms is essential. This necessity arises from the time-sensitive nature of foraging activities, where quick decision-making is pivotal to avoid potential harm.

## Potential Health Impacts:

- The repercussions of consuming toxic mushrooms are not to be taken lightly. From gastrointestinal distress to severe organ failure, the potential health impacts are alarming. This underscores the urgency in implementing a robust classification model that acts as a safeguard against the inadvertent ingestion of mushrooms with harmful consequences.

# Assumptions

---

## Dataset Representativeness:

- The dataset is believed to represent the overall mushroom population, ensuring observed patterns align with real-world scenarios.

## Variable Sufficiency for Edibility Predictions:

- Provided variables are deemed sufficient for accurate mushroom edibility predictions, encompassing key features crucial for distinction.

## Data Integrity and Reliability:

- Our dataset is assumed error-free, undergoing rigorous quality checks to ensure a foundation of reliable model outcomes.

## Stability of Variable Relationships:

- The relationship between variables and mushroom edibility remains stable over time, enhancing model generalizability.

# Data Description

## Our Data Treasure Trove

### Dataset Overview:

- Our dataset is a comprehensive compilation with 61069 records and 21 attributes.
- This abundance of data forms the bedrock for our predictive modeling endeavors.

### Feature Descriptions:

- There are 3 columns with numeric values, 2 with binary and 16 with categorical data.

### Target Variable: 'Class'

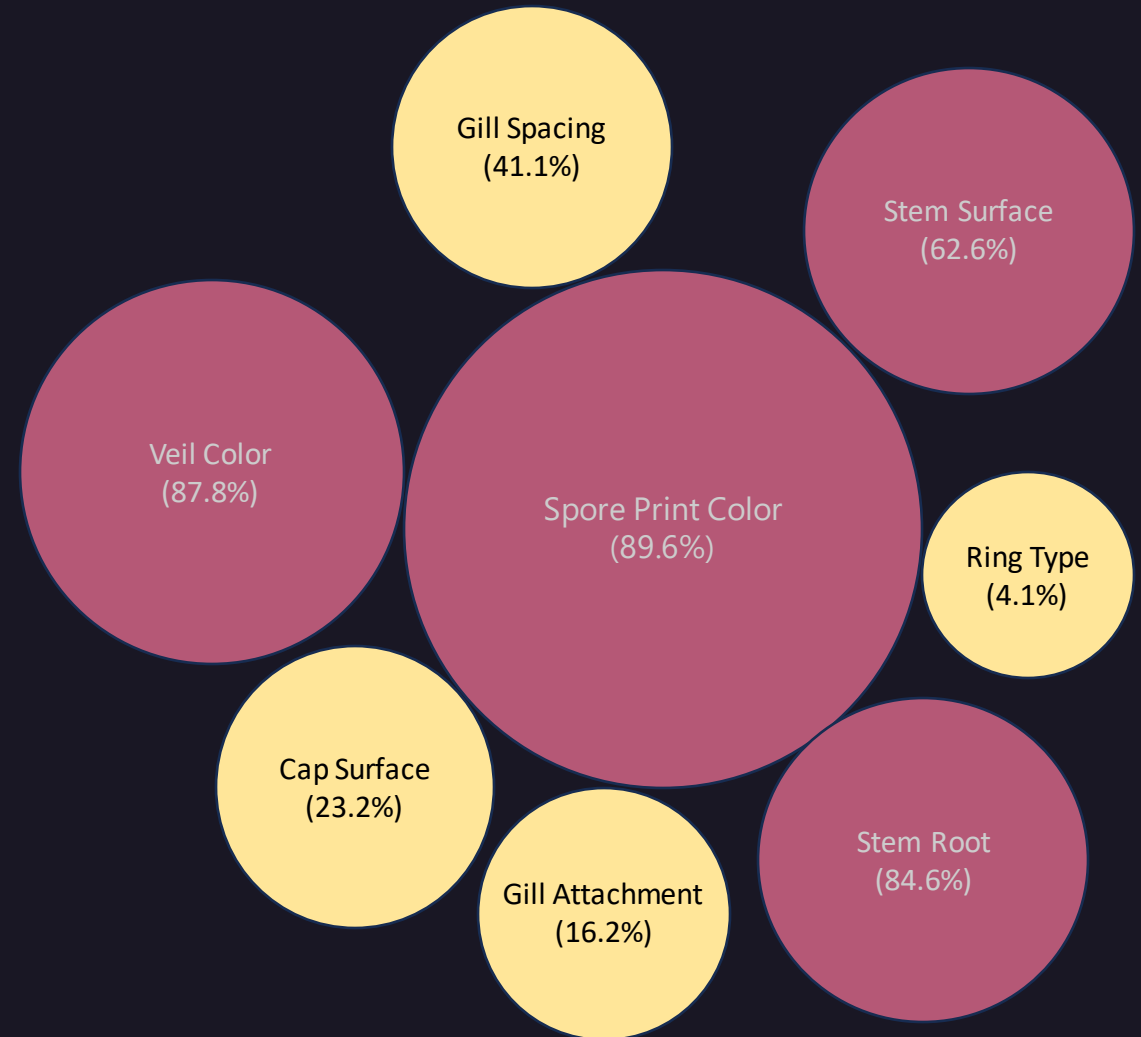
- At the heart of our classification model lies the 'Class' variable, a pivotal component indicating the edibility status of mushrooms.
- This binary variable distinguishes between two classes: Edible and Poisonous.
- Understanding the 'Class' variable is fundamental for our model's ability to accurately predict whether a mushroom is safe for consumption or poses potential harm.


Variable Name	Type	Description
cap-diameter	Numeric	Cap Diameter
cap-shape	Categorical	Cap Shape
cap-surface	Categorical	Cap Surface
cap-color	Categorical	Cap Color
bruise-or-bleed	Categorical	Bruise or Bleed (Yes/No)
gill-attachment	Categorical	Gill Attachment
gill-spacing	Categorical	Gill Spacing
gill-color	Categorical	Gill Color
stem-height	Numeric	Stem Height
stem-width	Numeric	Stem Width
stem-root	Categorical	Stem Root
stem-surface	Categorical	Stem Surface
stem-color	Categorical	Stem Color
veil-type	Categorical	Veil Type
veil-color	Categorical	Veil Color
has-ring	Categorical	Has Ring (Yes/No)
ring-type	Categorical	Ring Type
spore-print-color	Categorical	Spore Print Color
habitat	Categorical	Habitat
season	Categorical	Season
class	Categorical	The target variable indicating Mushroom edibility (Yes/No)
Physical Attribute	Structural Feature	Protective Structure
		Environ. Factor

# Data Cleaning

## Brewing the Data

- There are a lot of missing data, some columns have more than 80% of missing values.
- Columns having missing values are Spore Print Color, Veil Color, Stem Root, Stem Surface, Gill Spacing, Cap Surface, Gill Attachment, Ring Type, Out of above 8 variables, 4 variables have more than 60% missing data.
- There are no missing data in the binary and numeric columns.
- We shall deal with the missing data in the model creation part - because we want to avoid information leak.
- We shall drop redundant column 'veil-type', as it has constant value.





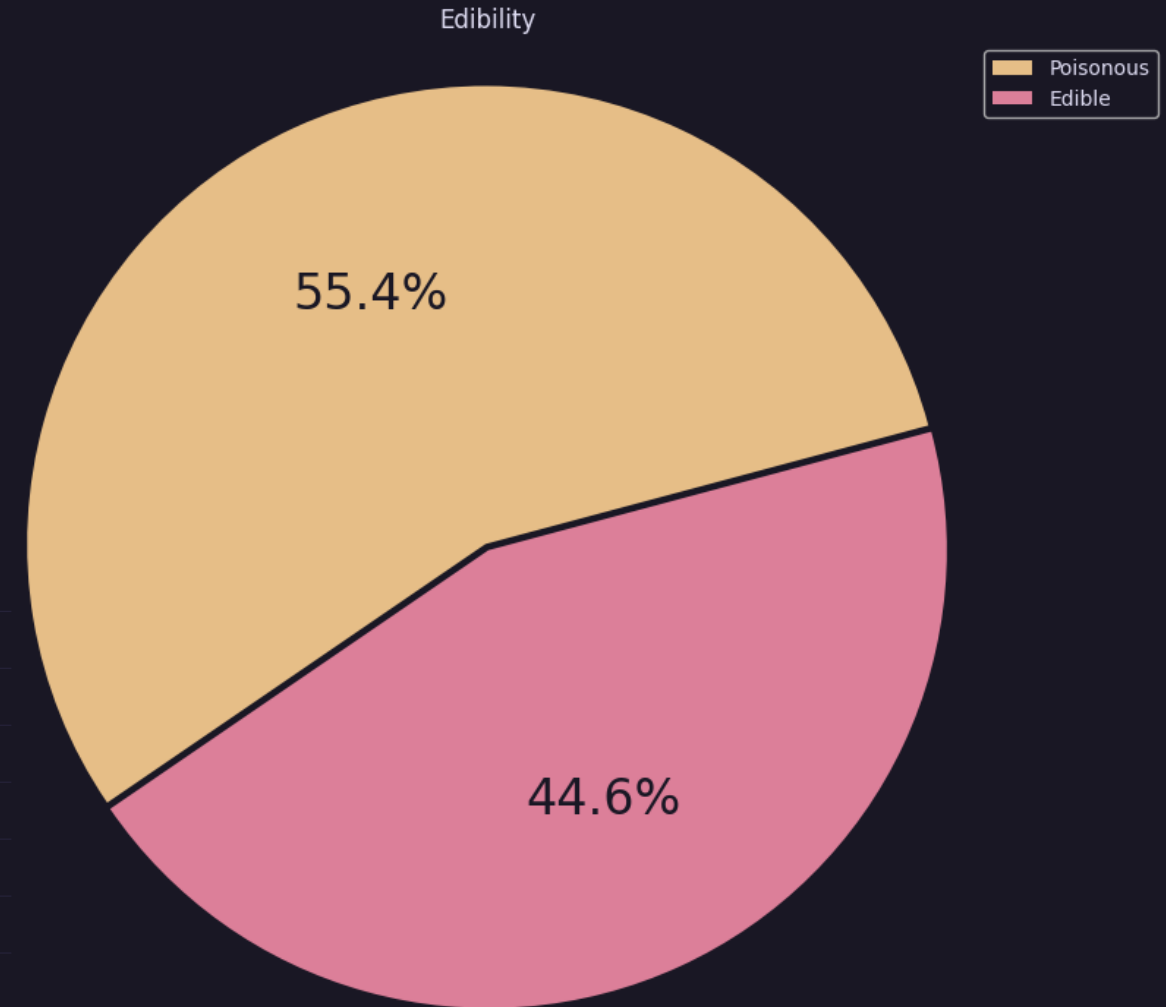
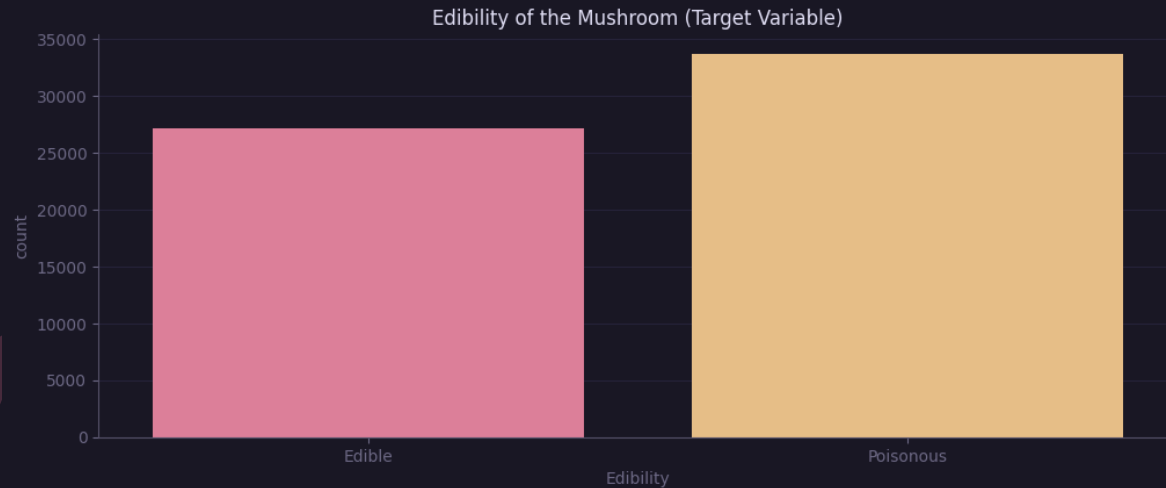
# Exploratory Data Analysis (EDA)



# EDA – Distribution Analysis

## Unearthing Insights

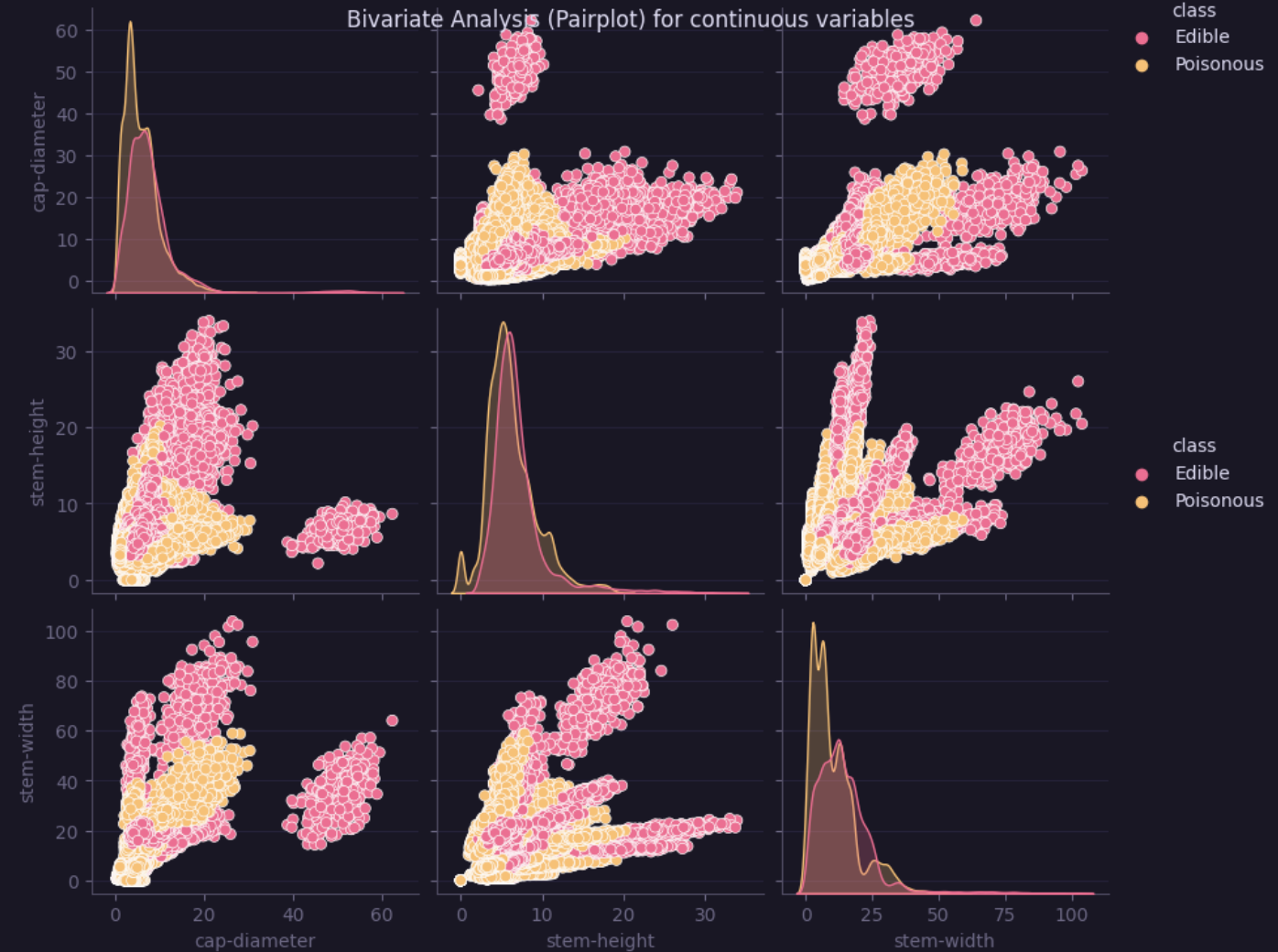
- Our target variable is balanced with (44.6%) records belonging to 'Edible' class while (55.4%) belonging to 'Poisonous' class.



# EDA – Distribution Analysis

## Unearthing Insights

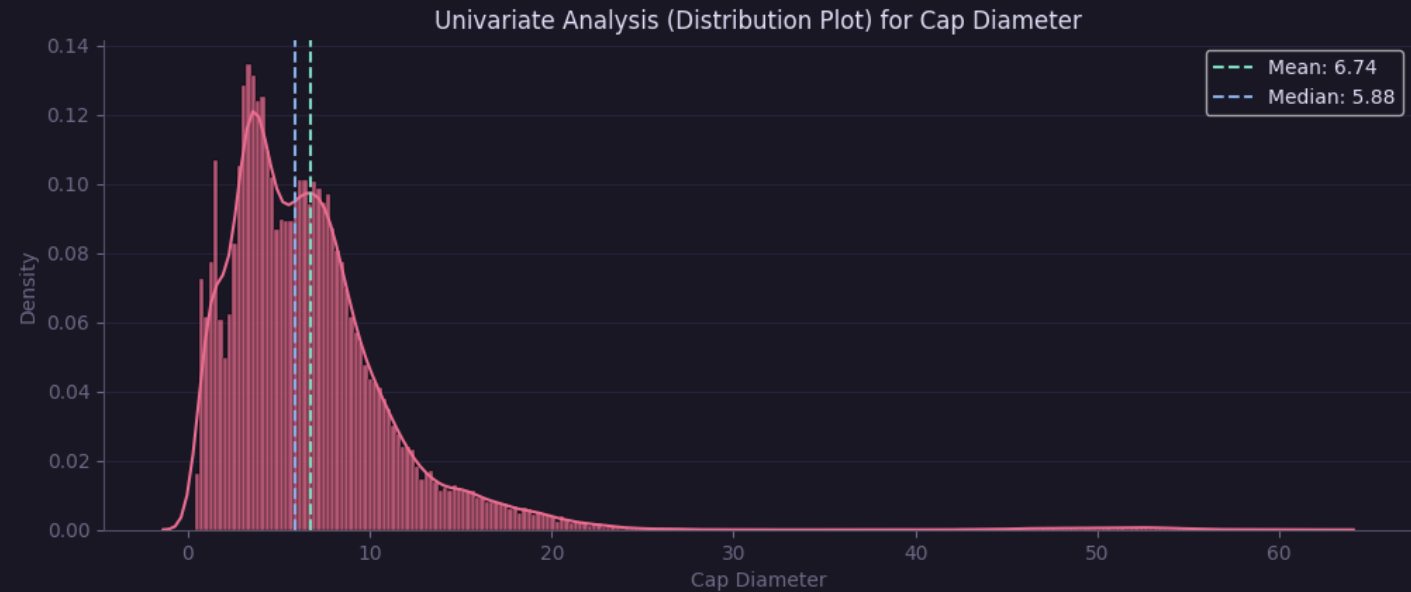
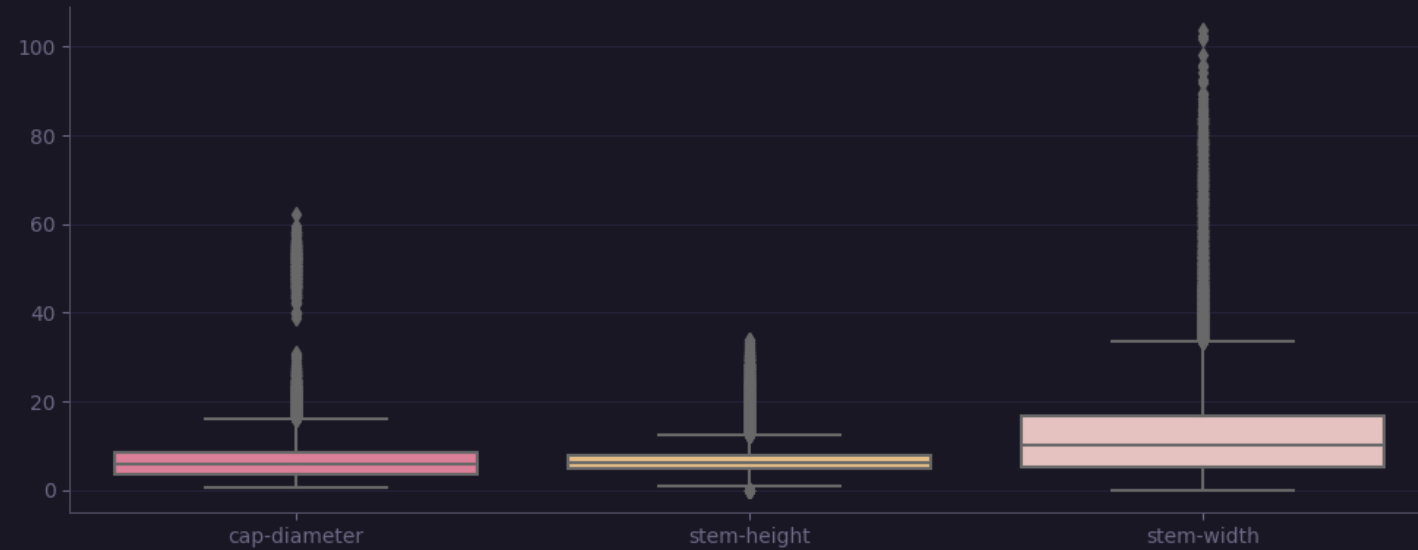
- All numeric columns have strong correlation with each other.
- Large Cap Diameter coupled with large Stem Height and/or Stem Width is a good indicator that a mushroom is edible.
- The mushrooms that are big are easy to tell apart. However, in the middle size range, there's a mix of both edible and poisonous mushrooms.



# EDA – Distribution Analysis

## Unearthing Insights

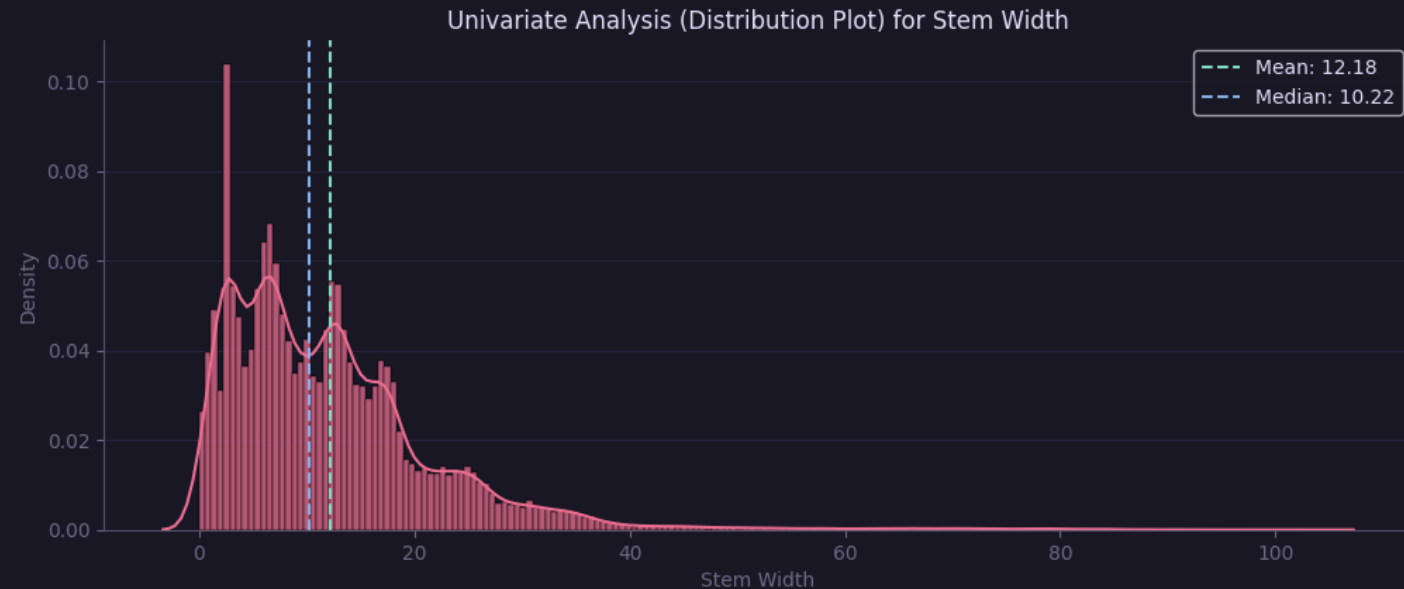
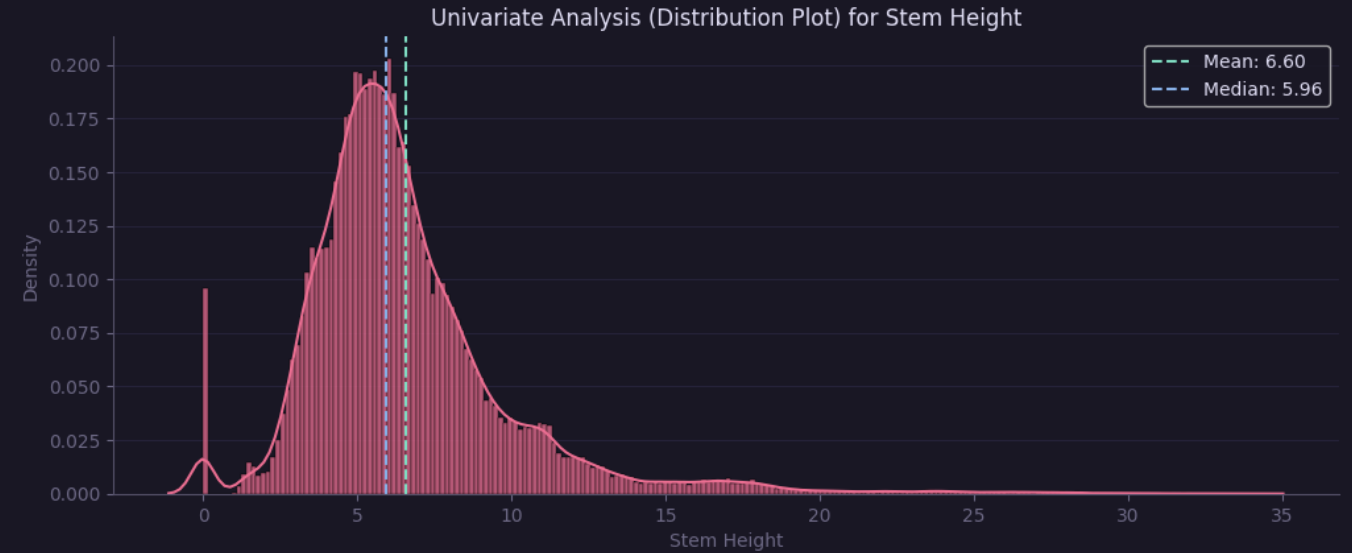
- All numeric columns have strong correlation with each other.
- Large Cap Diameter coupled with large Stem Height and/or Stem Width is a good indicator that a mushroom is edible.
- The mushrooms that are big are easy to tell apart. However, in the middle size range, there's a mix of both edible and poisonous mushrooms.
- The box plots reveal that this is a skewed dataset with many outliers. These outlier mushrooms tend to be edible.



# EDA – Distribution Analysis

## Unearthing Insights

- Large Cap Diameter coupled with large Stem Height and/or Stem Width is a good indicator that a mushroom is edible.
- The mushrooms that are big are easy to tell apart. However, in the middle size range, there's a mix of both edible and poisonous mushrooms.
- The box plots reveal that this is a skewed dataset with many outliers. These outlier mushrooms tend to be edible.

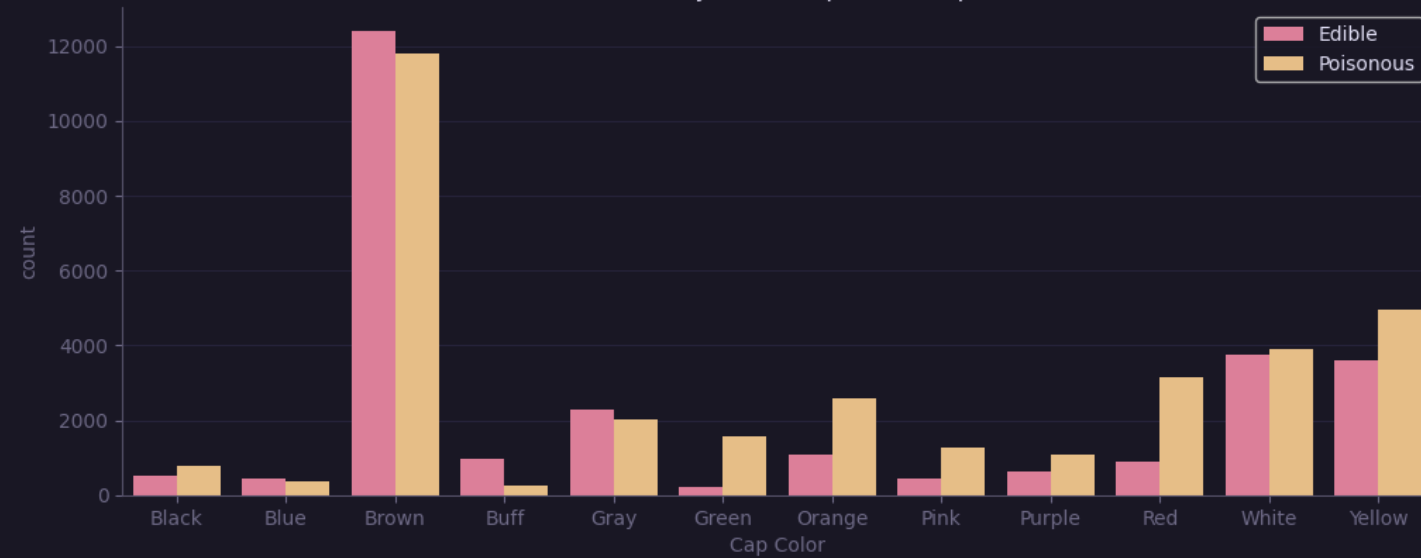


# EDA – Distribution Analysis

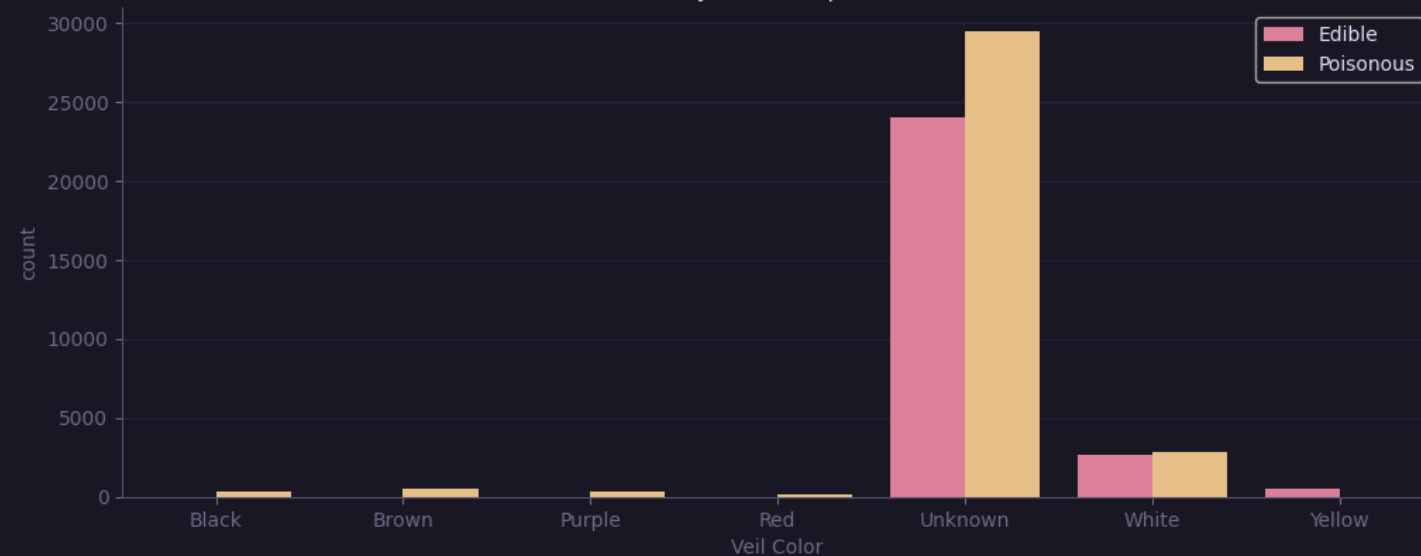
## Unearthing Insights

- Cap Color does not seem to be a good indicator of whether a mushroom is poisonous - except if the mushroom is red and orange. It's considerably more likely to be poisonous.
- Veil-color stands out as a feature that can conclusively identify number of edible mushrooms. Only white veils are inconclusive.

Univariate Analysis (Countplot) for Cap Color



Univariate Analysis (Countplot) for Veil Color

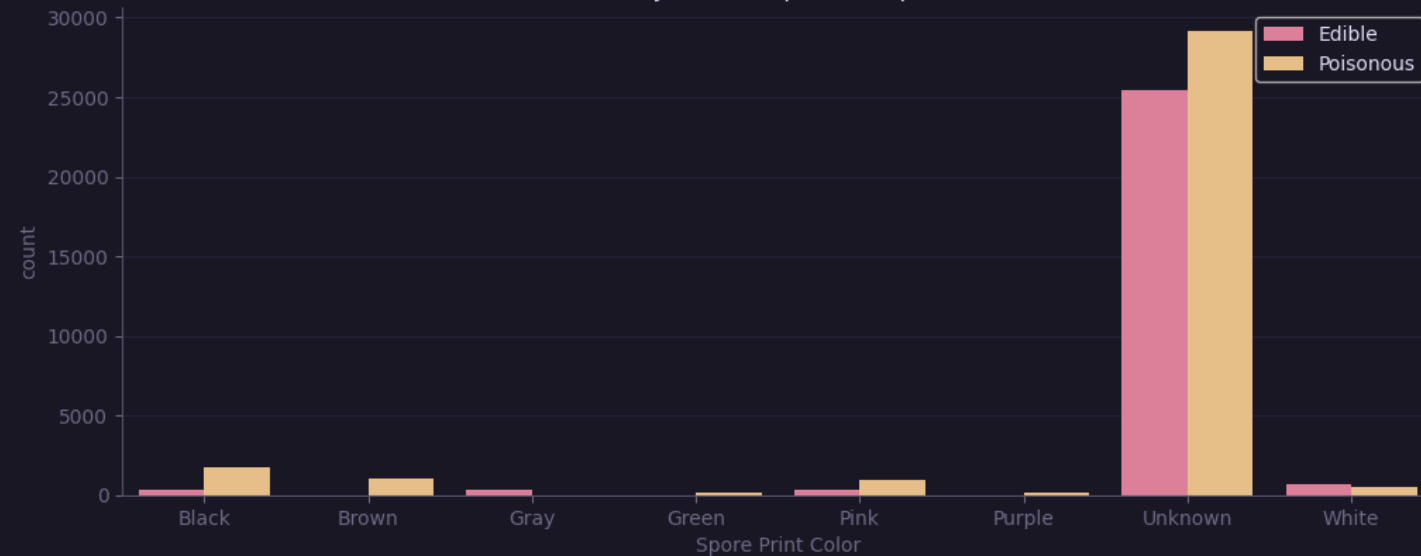


# EDA – Distribution Analysis

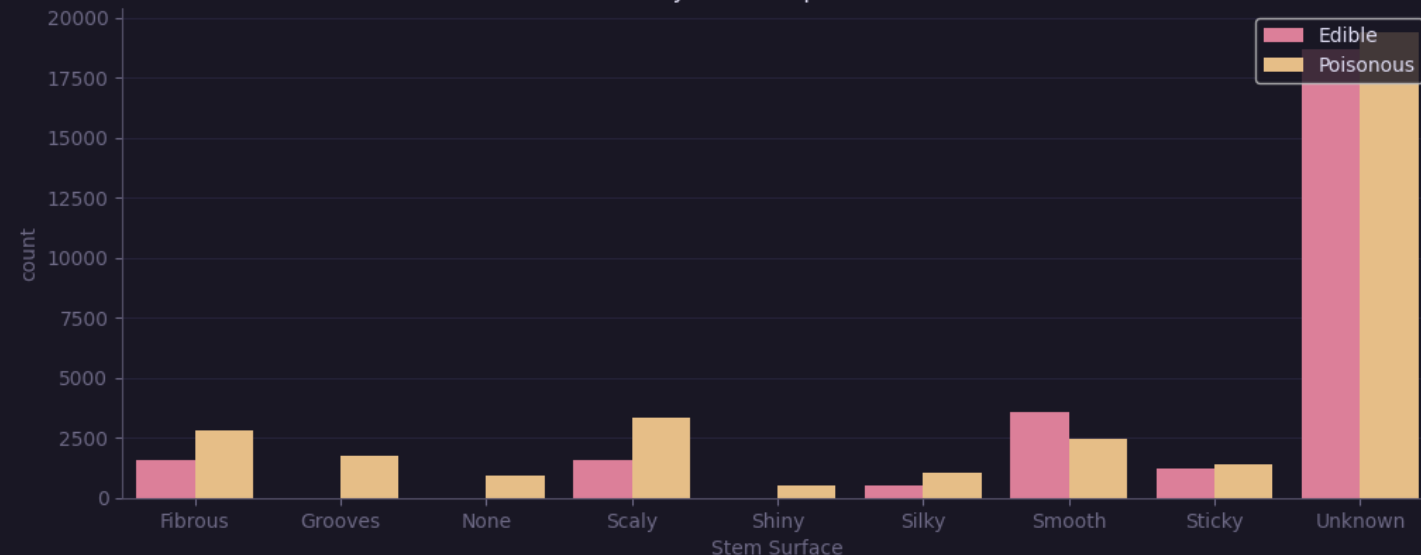
## Unearthing Insights

- Cap Color does not seem to be a good indicator of whether a mushroom is poisonous - except if the mushroom is red and orange. It's considerably more likely to be poisonous.
- Veil Color stands out as a feature that can conclusively identify number of edible mushrooms. Only white veils are inconclusive.
- Most likely, mushrooms with Gray Spore Print Color is edible while brown, green, purple are poisonous.
- Similarly, mushrooms with no stem or groovy and shiny stem surfaces are poisonous.

Univariate Analysis (Countplot) for Spore Print Color



Univariate Analysis (Countplot) for Stem Surface

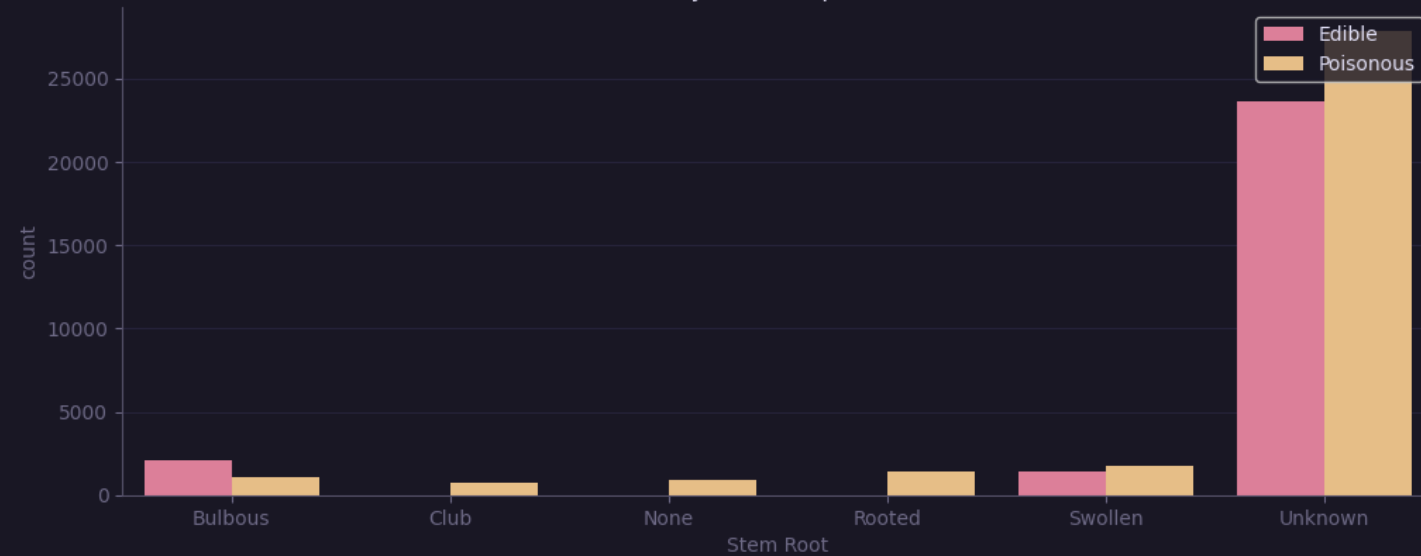


# EDA – Distribution Analysis

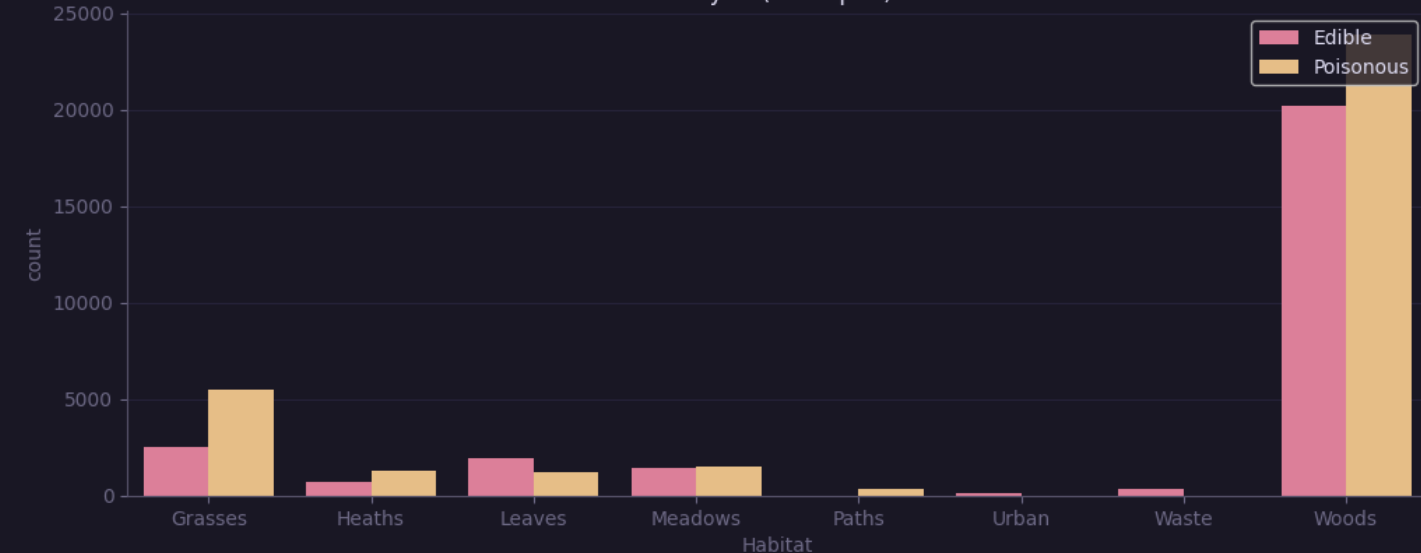
## Unearthing Insights

- Cap Color does not seem to be a good indicator of whether a mushroom is poisonous - except if the mushroom is red and orange. It's considerably more likely to be poisonous.
- Veil Color stands out as a feature that can conclusively identify number of edible mushrooms. Only white veils are inconclusive.
- Most likely, mushrooms with Gray Spore Print Color is edible while brown, green, purple are poisonous.
- Similarly, mushrooms with no stem or groovy and shiny stem surfaces are poisonous.
- Again, mushrooms with rooted, clubbed or no stem root are poisonous.
- Whereas mushrooms growing in the urban or waste habitats are edible. Notably, most mushrooms grow in woods.

Univariate Analysis (Countplot) for Stem Root



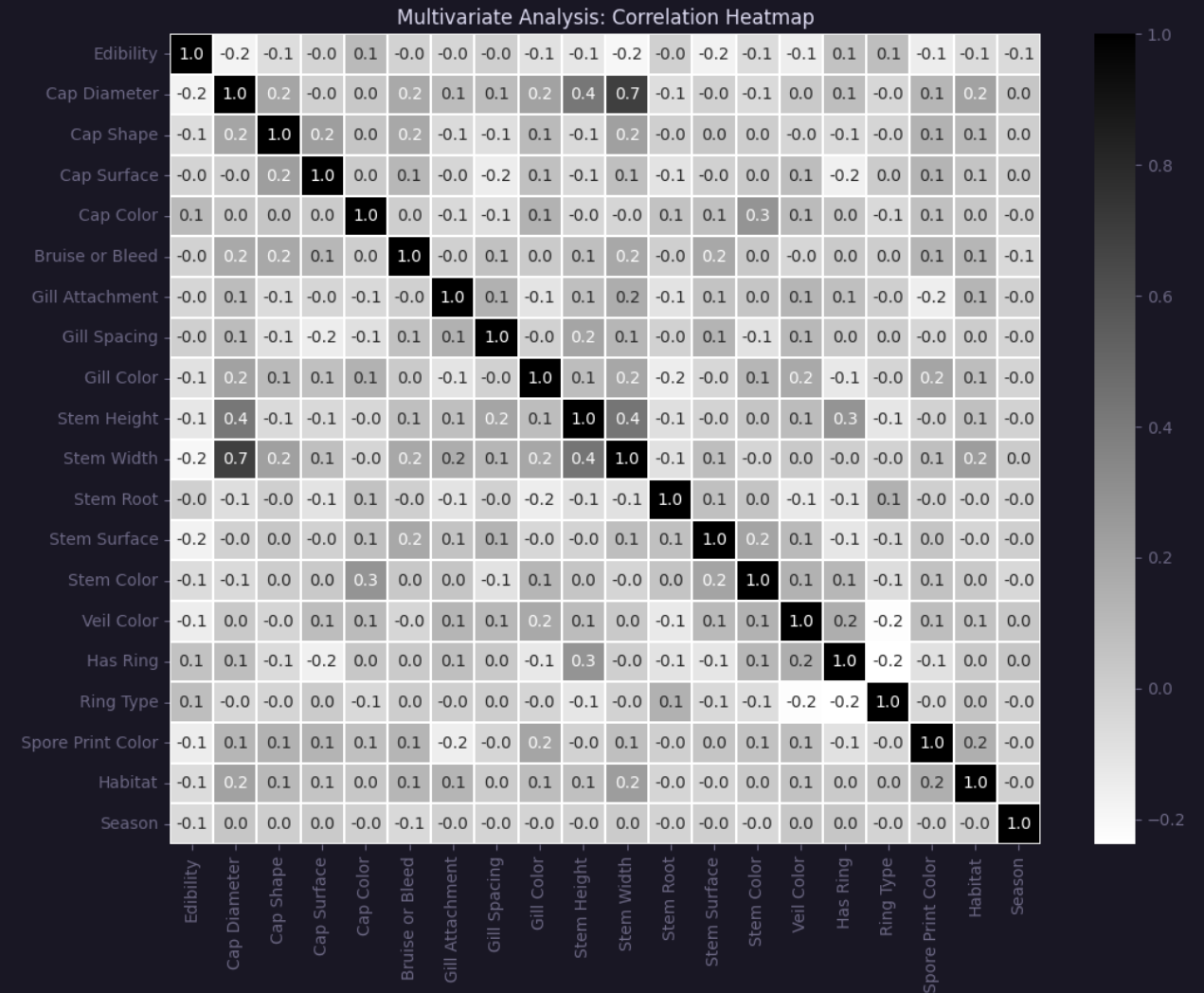
Univariate Analysis (Countplot) for Habitat



# EDA – Correlation Analysis

## Unearthing Insights

- All numeric columns have strong correlation with each other.





# Preprocessing & Baseline Modelling

## Missing Value Handling:

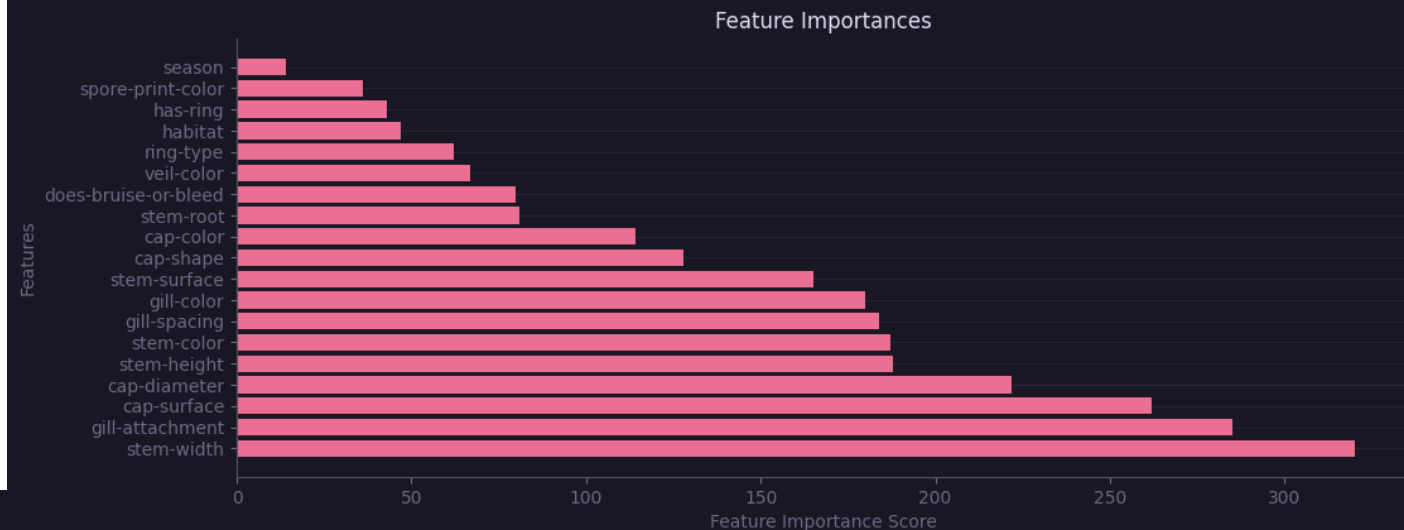
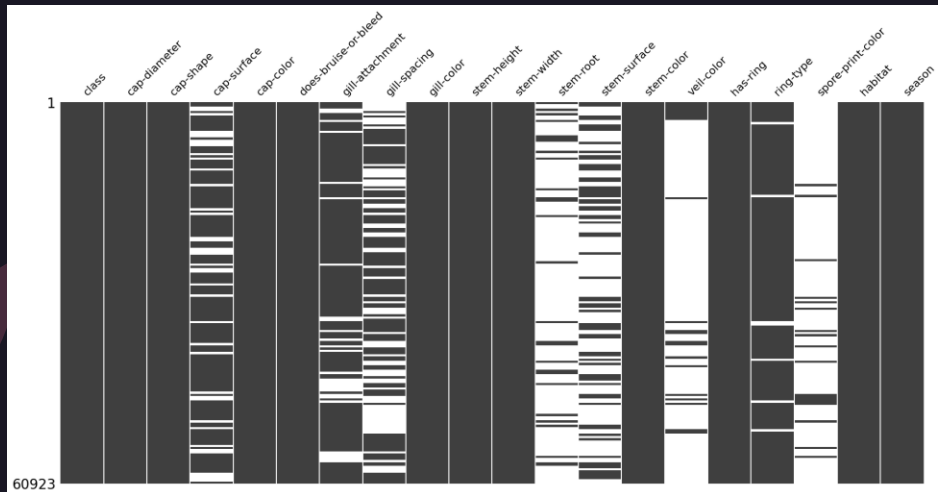
- Treating missing values as unknown helps maintain the integrity of the dataset, preventing distortion of patterns and relationships within the existing data. Assuming all missing values are unknown simplifies the handling of missing data, making it more practical and straightforward during the preprocessing phase.

## Columns for potential removal based on both feature importance and missing percentage:

- 'spore-print-color' (low importance: 36.0, high missing percentage: 89.62%) 'veil-color' (low importance: 67.0, high missing percentage: 87.83%) 'stem-root' (low importance: 81.0, high missing percentage: 84.59%) 'stem-surface', (low importance: 165.0, high missing percentage: 62.57%)

## Baseline Modeling with Dummy Classifier:

- We chose Dummy Classifier (Most Frequent Category) as our initial model with 55.9% accuracy.



# Model Building and Evaluation

## Scaling:

To ensure uniformity in feature scales, we applied StandardScaler to numeric columns. This preprocessing step is vital for algorithms sensitive to feature magnitudes.

## Classifier Initialization:

We initialized several classifiers, each with five-fold cross validation, to explore a variety of algorithmic approaches for our classification task. The chosen algorithms include XGBoost, Random Forest, Logistic Regression, K-Nearest Neighbors, Naïve Bayes, Decision Tree, and Support Vector Classifier (SVC).

Classifier	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)	ROC AUC(%)
Logistic Regression	78.01	74.87	76.60	75.72	86.27
Decision Tree	99.76	99.69	99.77	99.73	99.76
Naive Bayes	69.25	72.41	59.55	59.55	76.90
K-Nearest Neighbors	99.99	99.99	99.99	99.99	99.99
SVC	99.96	99.99	99.96	99.96	99.99
Random Forest	99.99	100	99.99	99.99	99.99
XGB	99.98	100	99.94	99.97	100

# Model Building and Evaluation

Selected Model (XGBoost/Random Forest):

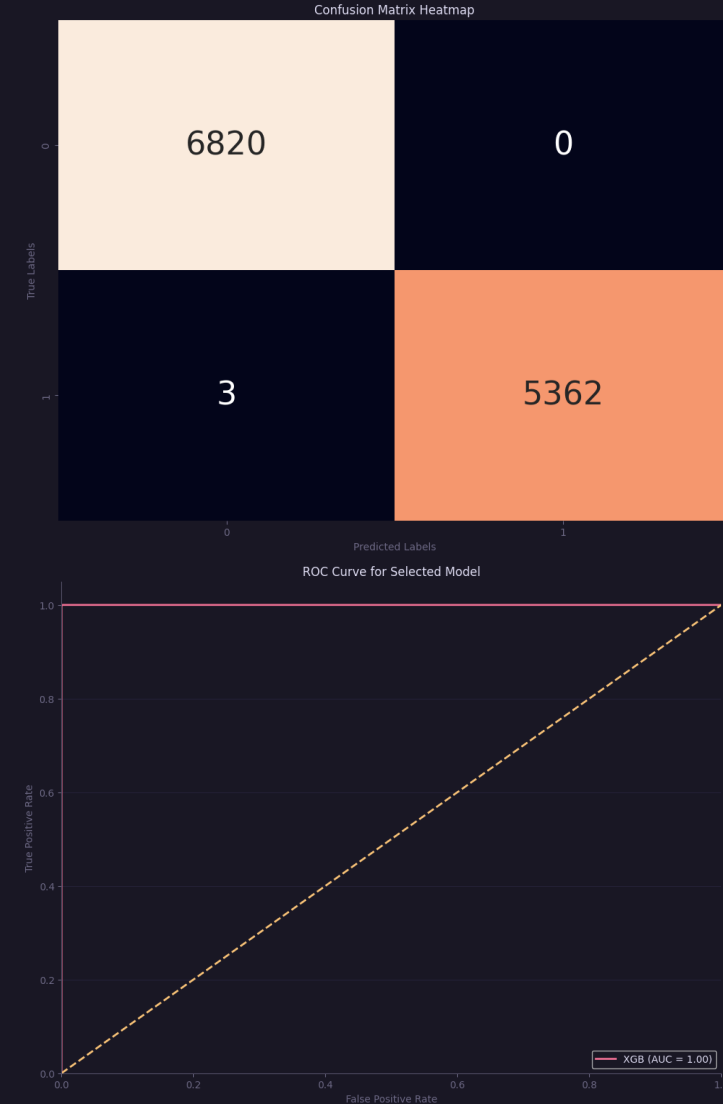
Based on the evaluation results, XGBoost and Random Forest emerged as the most promising model for our classification task, achieving the highest Precision, F1 Score, and ROC AUC among the classifiers.

Performance Metrics for Selected Model (XGBoost):

- Recall: 1
- Precision: 1
- F1 Score: 1
- ROC AUC: 1

Justification:

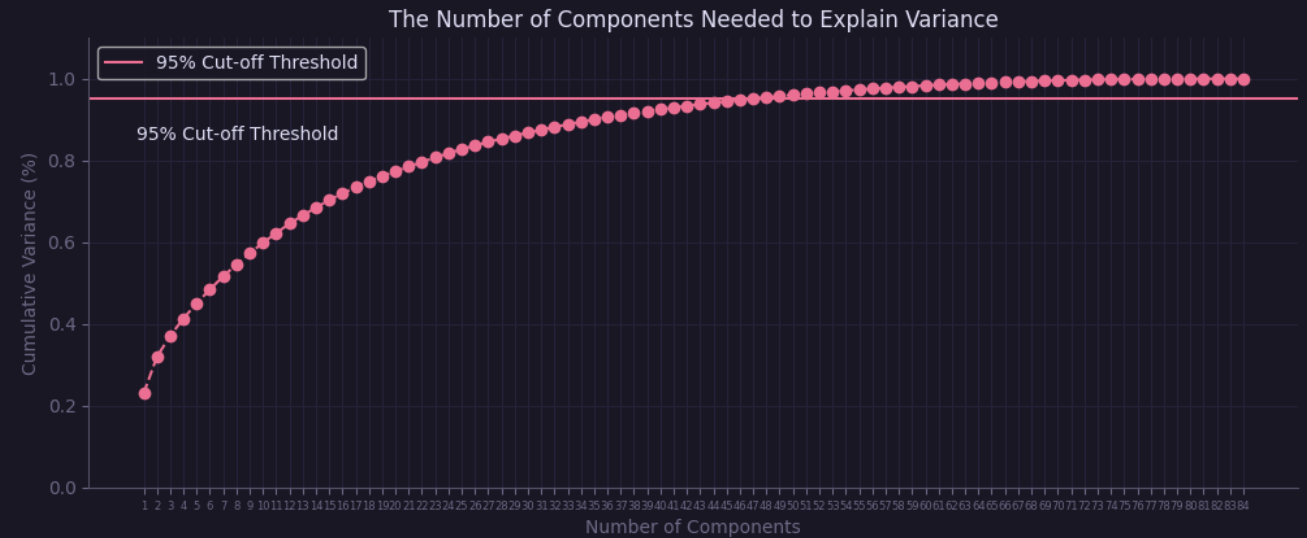
- Our choice of XGBoost as the selected model is validated by its superior performance in Precision, Recall, F1 Score, and ROC AUC, highlighting its potential to excel in classification tasks with categorical variables.



# Model Building and Evaluation

Unveiling the Path to Enhanced Predictive Power

- We employed PCA to identify the top components (41) to reduce dimensionality by focusing on the most relevant components.



"Thank you for joining us on this journey. Together, we can make a difference."